Dept. of Math. University of Oslo
Statistical Research Report No. 14
ISSN 0806–3842 December 2003

Statistical inference under symmetry.

Inge S. Helland, Department of Mathematics, University of Oslo,
P.O.Box 1053 Blindern. N-0316 Oslo, Norway

Abstract

We explore the consequences of adjoining a symmetry group to a statistical model. Group actions are first induced on the sample space, and then on the parameter space. It is argued that the right invariant measure induced by the group on the parameter space is a natural non-informative prior for the parameters of the model. The permissible sub-parameters are introduced, i.e., the subparameters upon which group actions can be defined. Equivariant estimators are similarly defined. Orbits of the group are defined on the sample space and on the parameter space; in particular the group action is called transitive when there is only one orbit. Credibility sets and confidence sets are shown (under right invariant prior and assuming transitivity on the parameter space) to be equal when defined by permissible sub-parameters and constructed from equivariant estimators. The effect of different choices of transformation group is illustrated by examples, and properties of the the orbits on the sample space and on the parameter space are discussed. It is argued that model reduction should be constrained to one or several orbits of the group. Using this and other natural criteria and concepts, among them concepts related to design of experiments under symmetry, leads to links towards chemometrical prediction methods and towards the foundation of quantum theory.

Key words and phrases: Confidence sets; Credibility sets; Group; Invariance; Invariant measure; Loss; Non-informative prior; Optimal estimator; Objective Bayes; Orbit; Partial least squares regression; Permissible sub-parameter; Pitman estimator; Quantum mechanics; Right invariant prior; Risk; Symmetry; Transitivity.

1 Introduction.

Corresponding to three different probability concepts (subjective, based on relative frequencies and based on symmetries), one may contemplate three different paths towards a foundation of statistical inference: the Bayesian, the frequentist and the symmetry-based. Of these, the last one has links to the other two, but may be

considered as the least developed of the three. Our aim here is partly to review its current status, and partly to try to contribute to its further development.

As recently pointed out by Breiman (2001), the theoretical statistical literature is to a large extent dominated by aspects of probability-based data models. That paper initiated an interesting discussion, but whatever attitude one should have here, it is highly relevant to ask whether also other structural elements can be of importance for statistical methodology, elements which in terms of their practical implications so far in our opinion has not been sufficiently focused upon in the literature. In this paper we will mainly study symmetry aspects by letting a group G act upon the parameter space and the sample space. This is of course not a new concept, but the tendency in the statistical literature has been to start with a model, and let the group be induced by this model. Here we want to argue for the group as an independent entity in addition to the model. The introduction of such a group will be shown to have several consequences for the analysis of data.

First, turn to the Bayesian approach. This literature here varies somewhat with respect to what should be meant by a non-informative prior (see the comprehensive review by Kass and Wasserman, 1996). When this term is used in situations with symmetry (in the simplest case location and/or scale symmetry), we claim that the non-informative prior may be taken to be the right invariant measure of the transformation group, a measure that can be defined in a natural way for any well-behaved group. (See definitions in Section 2 below.) We will give several arguments for the use of this prior later.

The Bayes estimators that are obtained using a right invariant measure as prior, will also have several other good properties, and could equally well have been arrived at by using these properties. When the loss function is invariant and the group action is transitive on the parameter space (see below), the estimators will be best equivariant (see below) under the group in question, and they will typically be minimax.

We will argue from several points of view that when one has prior symmetry information given by some fixed group, then this information should be made explicit by also using other quantities connected to the group, not only the invariant measure. This proposal is in conflict with the Bayesian view that all prior information should be expressible as a measure.

Finally, we discuss the concept of model reduction, and relate it to the orbits of the group as acting upon the parameter space. An orbit is the set of parameter values that can be arrived at from starting at one fixed value and letting the group elements transform this value. A transitive group has only one orbit. In general there is an optimal (best equivariant; Bayes using right invariant prior) estimator on each orbit. This implies that it is always natural to confine model reduction to be to one or several orbits of the group. First, in regression analysis this criterion

leads essentially to a model related to the chemometrician's partial least regression method, and the criterion seems to have the potential of being used in investigating several algorithmically defined methods. Secondly, and this may be important, by combining the criterion with some rather natural assumptions and including a design of experiment phase, it seems to be possible to find a new, non-formal way towards quantum mechanics. Thus in our view, quantum theory may in a natural way be regarded as a statistical theory.

A more detailed and technical version of parts of this paper, also containing some other examples, is given in Helland (2002a).

2 Group actions and statistical models.

Let a statistical model be defined as $\mathrm{P}^{\theta}(dx)$ with a sample space X and a parameter space Θ. Now introduce a group G into this setting, in the way proposed for instance by Fraser (1961, 1968). Such a group contains the set of symmetries that are natural to think of in the given situation. Briefly: Assume that the situation where data and parameters are transformed by some group element also may be considered to be an equally natural basis for inference as the original data and parameters. Then the collection of such elements constitute a natural group for the problem at hand. Changes of units of measurements constitute simple examples.

We start with a sample space X. A transformation group G is assumed to act on this space. Like other authors in this field, e.g., Dawid and Stone (1982), we will not focus much on topological and measurability questions. With a model as above, this also induces a class of transformations g on the parameter space by

$$\mathrm{P}^{\theta g}(A) = \mathrm{P}^{\theta}(Ag^{-1}). \tag{1}$$

This constitutes what mathematicians call a *homomorphism* between the two group actions: If g_1 and g_2 act on the sample space and then introduces similar actions on the parameter space by (1), then the product $g_1 g_2$ is mapped by (1) in a consistent way. This is why we can use the same symbols g, G for transformations on both spaces.

In other cases it is more natural to define the group actions directly on the parameter space. Even then any model must have the property that there also can be defined group actions on the sample space, and this in such a way that (1) holds.

A pure mathematician will think of G as an abstract set of elements with a multiplication table. As statisticians we are more interested in G as a transformation group: It introduces an action $\theta \mapsto \theta g$ on the parameter space, and similarly on the sample space by $x \mapsto xg$. For these transformation groups the usual group properties

are obvious: An associative multiplication $g_1 g_2$ is defined between all group elements, there is a unit $e \in G$ and an invers g^{-1}.

We will consistently write the group elements on the right in the transformations xg and θg. The main motivation for this is to make it easier to derive the right invariant prior in a natural way. Another motivation is that it may lessen the danger of confusing group transformation with multiplication.

In principle the statistical model implies very few restrictions on the choice of group: At this point only that the full spaces X and Θ should remain invariant under the transformations. Later we will require that the loss function shall be invariant under the group and that the right invariant measure of the group is a natural non-informative prior to use.

Nevertheless the choice of a transformation group may often be related to the selected statistical model; in particular to the choice of conditioning in the model. One case where this can be illustrated, is in connection to the choice of conditioning in multi-way contingency tables; see references in Helland (1995). Another case is the choice of random and fixed effects in analysis of variance; see Dawid (1988) and Helland (1998).

The fact that there exist cases where each of several groups in principle can be chosen in a given situation has been used (among others by Berger, 1985) as a counterargument against the group approach in general. Against this one can argue that that fact that the choice of a group may be difficult, in principle is not more strange than the wellknown problem that the choice of a model may be difficult, or that there may be several different choices of a loss function.

Here is an example where the same model class in a very natural way can be endowed with very different transformation groups according to the situation: Many statistics textbooks make a point of the fact that both regression and analysis of variance problems can be handled by similar linear models of the form $y = X\beta + e$. Nevertheless, most applied researchers feel strongly that there are essential differences between the two situations. From our point of view the difference is clear: In the regression situation it is natural to use the linear group $\beta \mapsto A\beta + a$ or some subgroup; in the case where all x-variables are measured in different units, it may be relevant to go to the subgroup defined by $\beta_j \mapsto k_j \beta_j; j = 1, ..., p$ for the regression components. On the other hand, in the analysis of variance situation the natural group is some permutation group, the choice of group depending upon the randomization used. (We will say more about this in Section 8 below.)

Simple groups in any data situation are the translation group defined by $xg = x + a$, the scale group $xg = bx$, where $b \neq 0$, and the translation- and scale-group given by $xg = bx + a$. One may also imagine extensions where a and/or b are given different values for data from different sources. Another common group is the rotation group in multivariate analysis.

Note that the concrete actual appearance of the group actions on the sample space and on the parameter space may be different. Here is an example: Let the multivariate data $(\mathcal{X}, y)$ have i.i.d. rows with x-covariance matrix Σ_{xx}, y-variance σ_y^2 and (x, y)-regression vector β. Let G be defined by rotations in the x-space: $(\mathcal{X}, y) \mapsto (\mathcal{X}Q, y)$ for orthogonal matrices Q. Then the induced transformations in the parameter space are $(\Sigma_{xx}, \beta, \sigma_y^2) \mapsto (Q'\Sigma_{xx}Q, Q'\beta, \sigma_y^2)$, apparently something different than the sample space group action.

Look again at the group of transformations acting on the sample space. For a given point $x_0 \in X$, let the *orbit* generated by this point be $x_0G = \{x : \ x = x_0g \text{ for some } g\}$. These orbits constitute equivalence classes in X; so we can always index the classes by some random variable a.

The group acting on the sample space is called *transitive* if there is only one orbit. Then each point x_1 can be transformed into every other point x_2 by some group element, and there is only one trivial, constant, value of a. If for every pair of points x_1 and x_2 there is not more than one group element g which transforms x_1 into x_2, we say that the group is *free*. This means that the group transforming a given element into itself is trivial; in general this group is called the *stability group*. Stability groups for different elements may b transformed in a simple way into each other in the transitive case.

If the group is both transitive and free, then one can pick one arbitrary basis point x_0 and write every element x in a unique way as $x = x_0g$. Thus in this case there is a one-to-one correspondence between the group and the sample space.

Similarly, for the group on the parameter space, we can have either of four basically different situations: Free or not free/ transitive or not transitive group.

Using (1) it is straightforward to prove the following: *The distribution of the orbit index a in the sample space depends only upon the parameter τ, the orbit index in the parameter space. In particular, if the parameter group is transitive, then the orbit index a is ancillary, i.e., has a distribution which is independent of any model parameter.*

3 Right invariant measure as a prior.

An important issue is that the choice of a symmetry group in a statistical setting also implies a natural choice of a non-informative prior distribution. First, it is well known (Nachbin, 1965) that every locally compact group possesses a socalled *right invariant Haar measure* ν on the group itself, that is, a measure with the property that $\nu(Dg) = \nu(D)$ when $g \in G$ and D is a set in the space of group elements. In the transitive case this measure is unique except for a multiplicative constant. When G is compact, ν can be taken to be a probability measure. In general

there is also a (dual) left invariant Haar measure μ (i.e., $\mu(gD) = \mu(D)$). This is equal to ν (except possibly for a multiplicative constant) for compact groups or if the group is commutative. In general one has $\mu(dg) = \Delta(g)\nu(dg)$ for the socalled *modulus* or modular function Δ, which also satisfies $\nu(gD) = (\Delta(g))^{-1}\nu(D)$ and $\Delta(g_1 g_2) = \Delta(g_1)\Delta(g_2)$.

Now turn to measures on the parameter space. If G is a group of transformations acting on this space, then our first recommendation concerning a non-informative prior on Θ is that it should be relatively invariant, that is, satisfy $\nu(\Gamma g) = \delta(g)\nu(\Gamma)$ for some multiplier (multiplicative function) δ, where Γ is an arbitrary set in the parameter space. In particular, if the multiplier is 1, we say that ν is a right invariant prior, so then $\nu(\Gamma g) = \nu(\Gamma)$. Existence of a relatively invariant measure ν can be shown to follow under general assumptions. In fact the existence of a right invariant measure on the parameter space follows under weak assumptions (the stabilty group in the parameter space should be compact; see Helland, 2002a) from the existence of the right Haar measure on the group itself. The connection is relatively simple: If for some fixed parameter value θ_0 the function β is defined by $\beta(g) = \theta_0 g$, then $\nu(E) = \nu_G(\beta^{-1}(E))$, where ν_G is the right Haar measure on the group. This can often be calculated using a suitable Jacobi-determinant. When Θ is non-compact, which usually is the case, ν will be an improper prior.

In certain cases it is obvious what the invariant measure will be. For example, for the translation group given by $\mu \mapsto \mu + a$ we get $\nu(d\mu) = d\mu$, while the rotation group for a p-dimensional vector μ has the natural rotation measure: Uniform distribution on the p-dimensional unit sphere. Some other relatively simple cases are: For the scale group $\sigma \mapsto b\sigma$ the invariant measure is $\nu(d\sigma) = d\sigma/\sigma$, while for the combined translation- and scale-group $(\mu, \sigma) \mapsto (b\mu + a, b\sigma)$ it is $\nu(d\mu, d\sigma) = d\mu d\sigma/\sigma$ (see, for instance Berger, 1985).

In other cases the construction of ν is not so simple. The following general rule (Berger, 1985) is useful: Assume that the transformation group G on the p-dimensional parameter space Θ can be considered as a subset of R^p with positive Lebesgue measure. If $J_g(h)$ is the Jacobi-determinant for the transformation of G given by $h \mapsto hg$, and if e is the unit element of the group G, then $\nu_G(dg)$ will be a measure with density

$$h(g) = \frac{1}{J_g(e)}.$$

One can give several good arguments for using the right invariant measure as a prior under symmetry, even in cases where this is an improper prior. A short review of these arguments follows; we refer to the literature for more details.

1. It is reasonable that the posterior measure at least should stay proportional if corresponding transformations are made of the sample space and the para-

meter space; we may let the constant of proportionality depend upon the group element. Simple arguments show that this - together with the invariance requirement on the model - implies that the prior must be relatively invariant.

2. A relatively invariant measure on a group G acting on Θ may induce a relatively invariant measure on Θ. This can always be done (under welldefined regularity conditions) in the right invariant case, otherwise it may or may not be possible.

3. The connection in 2) can be made through the function $\theta(\cdot)$ on G defined by $\theta(g) = \theta_0 g$, and the induced measure can be shown to always be independent of θ_0 in the right invariant case; not so in general in the other cases.

4. Recent results by Eaton and Sudderth (1993, 1995, 1998, 1999a,b) show that under very general conditions all other invariant inferences than those based upon right invariant priors are strongly inconsistent in a way which leads to uniformly inadmissible estimates, and which also is incoherent in a well-defined sense.

5. It is shown in Eaton and Sudderth (1999a) that Fisherian pivoting and the use of right invariant measure yield the same invariant predictive distribution under certain assumptions.

6. The Bayes estimators resulting from right invariant prior will also quite generally be best equivariant estimators (see below).

7. Bayes credibility regions will also be confidence regions under reasonable assumptions (Hora and Buehler, 1966; also, see later).

8. With a fixed group, and when inference is restricted to permissible parameters, the marginalization paradoxes of Dawid et al. (1973) are avoided (see Helland, 2002b) when right invariant prior is used.

9. Posterior distributions of invariant joint functions of parameters and data will under certain conditions have the 'correct' sampling distribution (see references in Dawid, 1983). For example, for i.i.d. $N(\mu, \sigma^2)$ data under the translation- and scale- group, when the right invariant prior (density $\propto \sigma^{-1}$) is used, the posterior distribution of the t-statistics will be a Student's distribution with $n-1$ degrees of freedom, but not so under the left invariant prior.

10. When proper priors converge to a right invariant measure, then the posteriors also converge as they should under weak assumptions (Stone, 1965).

11. This choice of prior leads to a close link to Fraser's structural inference (Dawid et al. 1973).

12. There are links to other non-informative priors (Kass and Wasserman, 1996),

As a balance against all these arguments, it should also be mentioned that there exist cases, admittedly rather extreme, (non-amenable groups; see Bondar and Milnes, 1981) where a right invariant prior may lead to an uniformly inadmissible estimator (Eaton and Sudderth, 1995).

4 Subparameters, inference and orbits.

Quite often in a statistical analysis, a subparameter, that is, some function of θ, is needed.

Requirement 1.
Inference should be limited to parametric functions that are permissible subparameters under the group G, that is, the parametric function $\eta(\cdot)$ should be such that $\eta(\theta_1) = \eta(\theta_2)$ implies $\eta(\theta_1 g) = \eta(\theta_2 g)$ for all $g \in G$.

This assumption then allows G to act as a transformation group on the range of $\eta(\cdot)$. The assumption, together with the use of right invariant prior, turns out to be enough to eliminate the marginalization paradoxes of Dawid et al. (1973), and also some similar inconsistencies; see Helland (2002a,b). It also implies under transitivity that credibility sets and confidence sets are equal with the same associated probability/ confidence level; see below.

If some given subparameter should not be permissible, one can always solve this in principle by going to a subgroup. In fact, one can easily show that there is a maximal subgroup of G with respect to which the subparameter is permissible. As a rule, this subgroup will not be transitive.

The above proposal may also seem to go some way towards resolving the difficulties in Fisher's fiducial inference as it is further developed in Fraser's (1968) structural inference; see also Fraser (1979). For instance, when (the multivariate) x is $\mathrm{N}_p(\mu, I)$, several authors (from Stein, 1959 on) have pointed out a discrepancy between the fiducial distribution of μ and that of $\mu'\mu$ (obtainable from the distribution of $x'x$). (See also a Bayesian discussion of the same example in Berger (1985), p. 230.) From the present point of view, an essential remark is that the two problems can be naturally related to two different groups, the group of translations and the group of rotations in $\mathcal{R}^p$. It is of importance then that the function $\mu \to \mu'\mu$ is *not* permissible with respect to the group of translations.

A standard concept in the statistical literature involving group invariance is the concept of *equivariant estimator* (see Lehmann and Casella, 1998), a concept which can be closely linked to that of a permissible parameter. Roughly speaking an estimator is equivariant if it transforms under the group in the same way as the

parameter to be estimated; a precise definition will be given in Section 5. The best equivariant estimator, which will be discussed in some detail below, will in general depend upon the group used (see examples given in Lehmann and Casella, 1998, and in Berger, 1985). Thus again the choice of group is crucial. In fact, the assumptions above imply that the formal Bayes estimators under right invariant prior are also the best equivariant estimators.

Another use of a specified group in at least some statistical inference problems (see also Fraser, 1968) is: One will usually condition the statistical analysis upon the orbits; at least if the orbit index has a distribution which is independent of the parameter (which it will when the parameter group is transitive; see the end of Section 2). Hence different choice of group may often mean different conditioning. The non-uniqueness of the ancillary statistics to condition upon is a well known problem in statistics. Specification of a group leads to a unique orbit index.

As already noted, the orbit index a in the sample space is ancillary if the parameter group is transitive. This is a very common situation, since, because one purpose of a statistical model is to condense information, the parameter space usually is 'less' than the sample space. A typical example is the case where the distribution of a set of observations $x_1, ..., x_n$ depend upon a location parameter μ and a scale parameter σ. Then it is natural to look at the translation- and scale-group $x_i \mapsto a + bx_i$ with the corresponding parameter group given by $\mu \mapsto a + b\mu$, $\sigma \mapsto b\sigma$. It is easy to see that the last transformation group is transitive, while the group in the sample space has orbits indexed by the socalled configuration, for instance given by

$$a = \{\frac{x_{i+1} - x_i}{x_2 - x_1}; i = 2, ..., n-1\}.$$

It is often claimed that all inference should be conditional, given such ancillary variables, in particular that the uncertainty shall be given as conditional, given a. If the observations are normally distributed, it follows from Basu's theorem that the mean $\overline{x}$ and the standard deviation s are independent of a, so it doesn't matter for the inference about (μ, σ) whether or not we condition with respect to a.

Here is one more statistical use of a given group, complementary to the one given above: The orbit index (a) in the sample space will be a maximal invariant under the group. Furthermore, the distribution of a only depends upon the maximal invariant (τ) in the parameter space. Useful inference on τ can therefore be performed using the marginal distribution of a, for instance computing maximum likelihood estimates from this distribution, not from the full sample distribution. The well known restricted maximum likelihood method for estimating variance components in mixed models can - as pointed out by McCullagh (1996) - be seen in this perspective; see also Helland (1999a).

So, as a conclusion, adding a group to the model specification is of interest, and does have consequences. On the other hand, the symmetry approach to statistical inference also implies difficulties, most notably the difficulty of choosing a group in a given case. In general, the symmetries expressed by the group should have some substantial basis in the concrete problem described.

5 Decision problems under symmetry.

It is well known that under quite general conditions (Berger, 1985; Lehmann and Casella, 1998) risk functions are constant functions of the parameters for decision rules which are invariant under some given group; in particular this holds for equivariant estimators, see below. We have been interested in finding (i) results in this direction that are constructive in the sense that explicit expressions can be given for the best equivariant estimators and (ii) the most general result of this kind. Partial results have among others been given by Fraser (1961), Stein (1965), Hora and Buehler (1966), Bondar (1972), Berger (1985) and Kariya (1989), but both our aims seem first to be achieved by the results by Eaton (1989) and Eaton and Sudderth (1999b); we will rely on the first of these sources here.

As before, let the sample space be X, the parameter space Θ and the model a family of probability measures $\{\mathrm{P}^\theta\}$ on X. In addition we need an action space - to be thought of as the space of values of a class of estimators. As before, we disregard measurability questions.

The group G of transformations is acting upon X, and also induces a group of transformations on Θ by (1). In this section we also assume that the measures P^θ are dominated by a fixed measure P on X, which we will assume is right invariant. We also assume that the densities

$$p(x|\theta) = \frac{d\mathrm{P}^\theta}{d\mathrm{P}}(x)$$

satisfy

$$p(xg|\theta g) = p(x|\theta)$$

for all $x,\ \theta,\ g$. From these properties we have that (1) holds, as desired.

Let $\eta(\cdot)$ be an permissible parameter as defined before, so that a group of transformations G is defined by $(\eta g)(\theta) = \eta(\theta g)$. An estimator $\hat{\eta}$ is then called *equivariant* if $\hat{\eta}(xg) = (\hat{\eta}g)(x)$ for all $g,\ x$. For more information on equivariant estimators, see Zacks (1971) and Bondesson (1982). If the a property like the above holds for an estimation problem (or more generally a decision problem), and if in addition the loss function L is invariant: $L(\hat{\eta}g, \theta g) = L(\hat{\eta}, \theta)$, we say that the decision problem is G-invariant.

In this and in next section we will assume that the parameter group action G is transitive, so that the risk is constant on the whole parameter space. It is easy to generalize to the intransitive case, however. As a weak technical requirement we will also assume that the group action G is proper (for a definition see Helland (2002a) and references there). We let ν_G be right Haar measure of the group G.

Theorem 1.
Let $\hat{\eta}(\cdot)$ be an equivariant estimator. Make the assumption above on G, L and on the model, and let ν_G be the right Haar measure on G. Then

$$\int_G L(\hat{\eta}(xg), \theta)p(xg|\theta)\nu_G(dg) = \int_G L(\hat{\eta}(x), \theta g)p(x|\theta g)\nu_G(dg) \tag{2}$$

For the proof we refer to Theorem 6.4 in Eaton (1989) specialized to the case with non-randomized actions.

The left-hand side of equation (2) is constant on the orbits of G in X. When averaged over the orbits according to the probability model, it gives the risk of the estimator. Since the group is assumed transitive on the parameter space, the right hand side is constant in θ, and this shows in particular that the risk is constant. But the strong point of this equation is that the righthand side can be used to find the best equivariant estimator explicitly, as demonstrated in several of the references mentioned above. Hora and Buehler (1966) interpreted the last integral as an expectation with respect to a 'fiducial' distribution, but also under certain conditions as an integral with respect to a posterior distribution. If L is a quadratic loss function, it is easy to see from (2) that the best equivariant estimator (Pitman estimator) can be interpreted formally as the Bayes estimator under right invariant prior.

Corollary 1.
Let $L(\hat{\eta}(x), \theta) = \|\hat{\eta}(x) - \eta(\theta)\|^2$, and assume that this loss function is invariant. Let the parameter group corresponding to G be transitive with right invariant measure ν. Then the best equivariant estimator for η is given by

$$\hat{\eta}(x) = \int \eta(\theta)\frac{p(x|\theta)}{p(x)}\nu(d\theta),$$

with $p(x) = \int p(x|\theta)\nu(d\theta)$. This estimator minimizes the conditional expected loss, given the orbit index for each orbit in the sample space.

Proof.
Use equation (2), expand and find the minimum of the quadratic form in $\hat{\eta}(x)$. Use the connection between the right Haar measure ν_G and the right invariant measure ν on Θ.

6 Credibility sets and confidence sets.

Essentially as in Berger (1985) an in other books on Bayesian statistics, define a $100(1-\alpha)\%$ credibility set as a set $C(x)$ in the parameter space whose posterior probability given data x is $1-\alpha$. We will consentrate on the non-informatiove right invariant prior ν, so that the posterior is $p(x|\theta)/p(x) \cdot \nu(d\theta)$, where $p(x) = \int p(x|\theta)\nu(d\theta)$. The credibility set is then defined by

$$\int_{C(x)} \frac{p(x|\theta)}{p(x)} \nu(d\theta) = 1-\alpha. \tag{3}$$

A confidence set $C(x)$ is also a set in the parameter space, depending upon data x, but the probability interpretation is completely different: $\mathrm{P}^\theta(\theta \in C(x)) = 1-\alpha$, where the probability is over the distribution of x. The link between the two concept, however, is easily found from Theorem 1, using $L(x,\theta) = I(\theta \in C(x))$.

Theorem 2.
Fix the orbit indices a and τ. Assume that the collection of sets $\{C(x)\}$ satisfies the transformation law $C(xg) = (Cg)(x)$ for all x and g. Then each $C(x)$ is a credibility set if and only if it is a confidence set, and the two probabilities associated with the sets are the same.

In fact, Theorem 1 gives a stronger statement: The corresponding conditional probabilities, conditioned upon the orbits in X, are equal. We will still assume that the group acts transitively on the parameter space.

Corollary 2.
Let $\eta(\theta)$ be a one-dimensional continuous permissible parametric function, and let $\hat{\eta}_1(x)$ and $\hat{\eta}_2(x)$ be two equivariant estimators. Define $C(x) = \{\theta : \hat{\eta}_1(x) \leq \eta(\theta) \leq \hat{\eta}_2(x)\}$. Then $C(x)$ is a credibility set and a confidence set with the same associated probability/ confidence level.

Proof.
Since the mapping g defined by $\eta(\theta g) = (\eta g)(\theta)$ is a continuous 1-1 mapping from a one-dimensional connected set onto another onedimensional set, it must preserve or reverse ordering. Without loss of generality, extend the definition of $C(x)$ to $\{\theta : \hat{\eta}_1(x) \leq \eta(\theta) \leq \hat{\eta}_2(x)\} \cup \{\theta : \hat{\eta}_2(x) \leq \eta(\theta) \leq \hat{\eta}_1(x)\}$. One of these components must be empty. So $(Cg)(x) = \{\theta : \hat{\eta}_1(x) \leq \eta(\theta g^{-1}) \leq \hat{\eta}_2(x)\} \cup \{\cdots\} = \{\theta : (\hat{\eta}_1 g)(x) \leq \eta(\theta) \leq (\hat{\eta}_2 g)(x)\} \cup \{\cdots\} = C(xg)$. Hence the result follows from Theorem 1.

A simple example is the following: Let $x_1, \ldots, x_n$ be i.i.d. normally distributed observations with sample mean $\bar{x}$ and sample standard deviation s. Let k be chosen so that the confidence statement $\bar{x} - ks \leq \mu \leq \bar{x} + ks$ has confidence coefficient $1 - \alpha$. Then the interval can also be given a definite probability interpretation: $1 - \alpha$ is also equal to the posterior probability of the interval when the prior is right invariant measure under the translation- and scale-group. Note that the action of G is transitive here. Similar examples can be found in a lot of cases where a non-informative prior is used in Bayesian analysis; see for instance Press (2003).

The results of this section also have immediate consequences for *confidence distributions*, an area which has been discussed recently by Schweder and Hjort (2003) as a frequentist alternative to Bayes posteriors. Briefly, if $[\eta \leq \eta_\beta(x)]$ is a one-sided confidence set for the parameter η with confidence coefficient β, and this is calculated for all β, the functional relation $F(\eta_\beta(x)) = \beta$ is equivalent to some $F(\eta_0) = \beta_0(x)$, which can be looked upon formally as giving a 'distribution' of η for fixed data, the confidence distribution of the parameter. A question of interest is when this is equal to a Bayesian posterior for some prior. The following immediate consequence of Corollary 2 gives a partial answer:

Corollary 3.
Assume that the statistical model is invariant under a group G, and that η is a one-dimensional continuous permissible parametric function of the model parameter θ. Assume that the group G is transitive on the parameter space. Then for fixed orbit index (ancillary) a under G the confidence distribution for η will be equal to the posterior distribution under right invariant prior.

Using Theorem 1 it is possible to generalize to the multiparameter case.

7 Examples. Orbits and model reduction.

We will not attempt any general thory of model reduction here, but the following remark seems rather obvious from the preceeding discussion: For each orbit of the group acting upon the parameter space, the Pitman estimator gives a good solution of any estimation problem within each orbit. Hence, if the purpose of a model reduction should be to be able to obtain more precise estimates from the model, there seems to be little reason to reduce the model within orbits. Using this argument, we will make the general recommendation to, once the group has been chosen, let any model reduction be to one or several orbits in the parameter space. Some examples which seem to support this recommendation, are:

- Look at two independent samples, $x_1, ..., x_m$, which are independently $N(\mu_1, \sigma_1^2)$ and $y_1, ..., y_n$, which are independently $N(\mu_2, \sigma_2^2)$. Use the translation- and scale-group given by $\mu_1 \mapsto a_1 + b\mu_1$, $\sigma_1 \mapsto b\sigma_1$, $\mu_2 \mapsto a_2 + b\mu_2$, $\sigma_2 \mapsto b\sigma_2$. (A common b must be used in order that $\mu_1 - \mu_2$ shall be permissible.) Then the orbits of the group in the parameter space are given by $\sigma_1/\sigma_2 = constant$. A very common model reduction is given by $\sigma_1 = \sigma_2$.

- Consider a two way analysis of variance with expectations $\mu + \alpha_i + \beta_j + \gamma_{ij}$, and a group generated by all permutations of the index i and by all permutations of the index j. Then an obvious reduced model is given by the orbit where the expectation is $\mu + \alpha_i + \beta_j$.

- In a multiple regression it is not uncommon that all explanatory variables x_j are measured in different units. Then a natural group in the sample space (permitting oneselves to include the covariates in this space) is given by separate scale changes $x_j \mapsto k_j x_j$ $(j = 1, 2, ...)$. This induces a similar group on the regression parameters β_j, and all orbits in the parameter space are given by putting some of the β_j's equal to 0. These reduced models are well-known from many applications of regression analysis, and criteria like C_p or AIC are are used to discriminate between them.

- Assume that you for some sample start by modelling it using an arbitrary non-parametric model. Let the usual translation- and scale-group act on this large parameter space. Then one orbit is given by the $N(\mu, \sigma^2)$ distribution, a not uncommon model reduction.

- Consider the rotation group for a multivariate data set. This induces the transformations $\Sigma \mapsto Q'\Sigma Q$ for the covariance matrix Σ, where Q is some orthogonal matrix. This is an extremely non-transitive group with orbits equal to every set of eigenvalues of Σ, counting multiple eigenvalues with their multiplicity. It is difficult to imagine a situation where it is of interest to take a single orbit as a reduced model, but sets of orbits can make interesting reduced models, say those where the number of different eigenvalues is some fixed number, or those where the 5 smallest eigenvalues are equal.

8 Design of experiments situations.

Consider a set Z of potential experimental units for some experiment; this set can be finite or infinite, one may even consider an uncountable number of units. For each given $z \in Z$, let y_z be some potential response variable, and let μ_z be the expectation of y_z for the case where no treatment is introduced. One may also have a set T of potential treatments which can be applied to each unit. Let μ_{tz} be the expectation of y_z, given z, when treatment t is applied to z, and define $\theta_{tz} = \mu_{tz} - \mu_z$. Assume for simplicity that the y_z's are independent with a variance σ^2. Let η_z denote other

parameters connected to the unit z.

In this situation it is natural to call $\phi = (\{\mu_z, \eta_z; z \in Z\}, \{\theta_{tz}; t \in T, z \in Z\}, \sigma^2)$ a total parameter for the system and $\Phi = \{\phi\}$ the total parameter space. This terminology is consistent with the one we will use in our approach to quantum mechanics below. Note that ϕ of course is not estimable in any conceivable experiment; nevertheless it is a useful conceptual quantity.

Let G be a transformation group defined on Z. This will induce a group on Φ. In other cases, larger groups on Φ may be of interest, but in the case of designed experiments it is permutation of the experimental units which is the important issue.

Now for the experiment itself select a finite subset Z_0 of Z. We will assume at the outset that G is so large that the full permutation group G_0 on Z_0 is a subgroup of G.

We will also assume that Z_0 is selected from Z by some random mechanism with the property that $\theta_t = \mathrm{E}(\theta_{tz}|t)$, expectation over this selection mechanism, is independent of the selected z. Then we will have for a given selected unit $z \in Z_0$ that

$$\mathrm{E}(y_z|t) = \mu_z + \theta_t.$$

This is one way to express the well known unit/treatment additivity, which is considered by Bailey (1981, 1991) and others to be crucial for having a consistent approach to the design of experiment.

From this point on Bailey (1981) introduces an eight-stage experimental design theory, and this theory is developed further in Bailey (1991). We will only mention very briefly a few main points of this theory, refering to these and related papers for details. Note that Bailey (2003) seems to give a relatively full account of the field of experimental design, including the many important practical aspects of this area.

Block structure is an important aspect of experimental design theory: Similar units are taken together in one block to enhance efficiency. This topic has many important facets, like Latin squares, split plot blocking, incomplete blocks and so on. From a group theoretical point of view, the main point is that the block structure determines the group used for randomization: For a selected experiment $\mathcal{E}^a$, use for randomisazion the largest subgroup G^a of G_0 which respects the block structure of that experiment: If the units z_1 and z_2 are in the same block, then $z_1 g$ and $z_2 g$ should be in the same block for all $g \in G^a$. The unit(names) are then randomized according to this group. This randomization is also crucial for the allocation of treatments.

Assuming that G^a is transitive, Bailey (1991) proves the following: After randomization, y_z (overusing this symbol slightly) has an expectation which only depends upon the treatment $t(z)$ given to z, and a covariance matrix C satisfying

$$C(z_1, z_2) = C(z_1 g, z_2 g), \tag{4}$$

for $z_1, z_2 \in Z_0$ and $g \in G^a$. Using this, Bailey (1991) introduces the *strata*, which are the eigenspaces of C, and which also are invariant spaces under the group G. The important practical point is that these give the lines of the (null) analysis of variance for the experiment, both in simple and in complicated cases.

9 Chemometrical prediction methods.

(This example is discussed in more detail in Helland, 1990, 2000, 2001.)

There exist several regression methods for collinear data, most of them have been derived from ad hoc considerations. Looking at a method like principal component regression, for instance, it is clear that any theoretical foundation of such a method must depend upon more than the conditional distribution of y, given x, specifically also upon the distribution in x-space. The same applies more generally. So in order to initiate some general theory connected to such methods, consider a p-dimensional x-distribution and the corresponding $(p+1)$-dimensional joint (x, y)-distribution with expectation $(\mu_x', \mu_y)'$ and covariance matrix

$$\begin{pmatrix} \Sigma_{xx} & \sigma_{xy} \\ \sigma_{xy}' & \sigma_y^2 \end{pmatrix}. \tag{5}$$

The parameter of interest here is the regression vector $\beta = \Sigma_{xx}^{-1}\sigma_{xy}$, and this is permissible under most groups of interest, in particular under the linear group $(x, y) \to (Ax + c, y)$ and all subgroups. Now inspect the estimated regression vectors $\hat{\beta}$ resulting from known regression methods like principal component regression, ridge regression, latent root regression and so on. None of these are equivariant under the full linear group, but they are all equivariant under the orthogonal group defined by $(x, y) \to (Qx, y)$ $(Q'Q = I)$, where we in particular get $\beta \to Q'\beta$. Hence it is of definitive interest to try to find a regression method which in some sense is optimal under this orthogonal group.

From prediction error consideration, it is easy to see that the reasonable loss function is

$$L(\hat{\beta}, \theta) = (\hat{\beta} - \beta)'\Sigma_{xx}(\hat{\beta} - \beta).$$

This is invariant under the orthogonal group, which gives $\Sigma_{xx} \to Q'\Sigma_{xx}Q$.

Now a complicating fact about the rotation group is that it is not transitive on the parameter space. Hence there is one Pitman estimator for each orbit of the group. These orbits are determined by the numbers and dimensions of the eigenspaces of Σ; the eigenvalues λ_k of Σ; the the norms γ_k of the components of β along the different eigenspaces and the values of σ^2.

Using the explicit form of the Pitman estimator it was argued in Helland (2001) that its most parameter dependence was only upon the number m of values k such that $\gamma_k \neq 0$.

Thus the most relevant model reduction seems to be through a fixation of m. This possibility immediately provides a link to the population model of a regression method developed in the last decades by chemometricians, namely partial least squares regression. This method was developed heuristically, and is still mainly presented as an algorithm. One possible conclusion from arguments like the one above, is that the set of models corresponding to partial least squares seems sensible, especially since the choice of model, being indexed by a one-dimensional parameter, is relatively easy, but that the estimation under the model probably can be improved.

10 The statistical approach to quantum mechanics.

The discussion below is developed considerably further in Helland, 2003a,b, where it is argued that quantum theory basically can be regarded as a statistical theory. Historicly, the distance between these two areas has been large. Quantum mechanics was developed by heuristic reasoning from physical observations by a considerable number of people in the beginning of the last century, a development which culminated in the formal theory put forward by von Neumann (1932). Since then, the basical theory has remained the same, but much further development has been made, and so-called paradoxes within the theory are still being vigorously discussed in the physical literature.

The development in Helland (2003a) starts by assuming a total parameter space Φ and a group G defined on this space. Then an experiment $\mathcal{E}^a$ is selected, having parametre $\theta^a = \theta^a(\phi)$ and the maximal group G^a such that θ^a is permissible. Further model reduction is done taking into account the orbits of G^a.

For the rest of the development towards quantum mechanics, the reader is refered to Helland (2003a,b). Our main point is that in our view, quantum theory and statistical theory can essentially be developed from the same setting if the appropriate concepts are used for such a starting point.

In the long run this joining of two disciplines using a similar language *may* result in a situation where researchers from different cultures may learn from each other. As I see it, this can be particularly useful because both disciplines, in addition to being connected to a formal language, also each are closely tied to an empirical basis.

11 Concluding remarks.

It might be appropriate here to cite from Efron (1998): 'A widely acceptable objective Bayes theory, which fiducial inference was intended to be, would be of immense theoretical and practical importance.'

One purpose of this paper has been to clarify some problems connected to situations for which a fairly acceptable objective Bayes theory - a theory of optimal inference under invariance - is available. The price paid for this coherent theory seems to be twofold: One has to fix a symmetry group for the problem at hand, and inference must be limited to parameters that are permissible under this group.

In many simple cases the choice of group is rather obvious, but it seems to be a challenge to find good, general rules for choosing the group in more complicated cases. Expressing lack of information in symmetry terms might be one way to proceed. Sometimes several groups lead to the same solutions.

Another question is whether the class of allowable parametric functions can be extended in any useful general way beyond the permissible ones. As illustrated in several cases above, however, this class can often be made rich enough for practical purposes by a suitable choice of group. As a general point, it must be more important to avoid incoherences than to be able to make inference on every possible parametric function.

Of course, then, at last: There are situations where it is not natural to choose a symmetry group at all before doing statistical inference, and there are other cases where it does not help much to choose a group at all even if this choice is made in a reasonable way. (Two such examples - related, and both attributed to C. Stein - can be found in Lehmann (1959) p. 231 and Berger (1985), p. 420.) A final open question is therefore if any of the ideas in this paper can be generalized also to certain specific situations with which it is difficult or useless to associate such a strong structure as a symmetry group.

References.

Bailey, R.A. (1981). A unified approach to design of experiments. *J. R. Statist. Soc. A* **144**, 214-223.

Bailey, R.A. (1991). Strata for randomized experiments. *J. R. Statist. Soc. B* **53**, 27-78.

Bailey, R.A. (2003). *Design of Comparative Experiments.* Book to appear. Draft available at http://www.maths.qmw.ac.uk/ rab/DOEbook/

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, Berlin.

Bondar, J.V. (1972). Structural distributions without exact transitivity. *Ann. Math. Statist.* **43**, 326-339.

Bondar, J.V. and P. Milnes (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **57**, 103-128.

Bondesson, L. (1982). Equivariant estimators. In: Kotz, S., N.L. Johnson and C.B. Read [Ed.] *Encyclopedia of Statistical Sciences.* Wiley, New York.

Breiman, L. (2001). Statistical modelling: The two cultures. *Statistical Science* **16**, 199-231.

Dawid, A.P. (1983). Invariant prior distributions. In: Kotz, S., N.L. Johnson and C.B. Read: *Encyclopedia of Statistical Sciences.* John Wiley, New York.

Dawid, A.P. (1988). Symmetry models and hypotheses for structural data layouts. *J. R. Statist. Soc. B* **50**, 1-34. *Ann. Statist.* **10**, 1054-1067.

Dawid, A.P.. M. Stone and J.V. Zidek (1973). Marginalization paradoxes in Bayesian and structural inference, *J. R. Statist. Soc. B* **35**, 189-233.

Eaton, M.L. (1989). *Group Invariance Applications in Statistics.* Institute of Mathematical Statistics and American Statistical Association, Hayward, California.

Eaton, M.L. and W.D. Sudderth (1993). Prediction in a multivariate setting: Coherence and incoherence. *Sankhya* A **55**, 481-493.

Eaton, M.L. and W.D. Sudderth (1995). The formal posterior of a standard flat prior in MANOVA is incoherent. *J. Italian Statist. Soc.* **2**, 251-270.

Eaton, M.L. and W.D. Sudderth (1998). A new predictive distribution for normal multivariate linear models. *Sankhya* A **60**, 363-382.

Eaton, M.L. and W.D. Sudderth (1999a). Consistency and strong inconsistncy of group-invariant predictive inferencees. *Bernoulli* **5**, 833-854.

Eaton, M.L. and W.D. Sudderth (1999b). Group invariant inference and right Haar measures. To appear in *J. Stat. Planning and Infer.*

Efron, B. (1998). R.A. Fisher in the 21st century. *Statistical Science* **13**, 95-122.

Fraser, D.A.S. (1961). The fiducial method and invariance. *Biometrika* **48**, 261-280.

Fraser, D.A.S. (1968). *The Structure of Inference.* Wiley, New York.

Fraser, D.A.S. (1979). *Inference and Linear models.* McGraw-Hill, New York.

Helland, I.S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97-114.

Helland, I.S. (1995). Simple counterexamples against the conditionality principle. *Am. Statistician* **49**, 351-356; discussion: **50**, 382-386.

Helland, I.S. (1998). A population approach to analysis of variance models. *Scand. J. Statistics* **25**, 3-15.

Helland, I.S. (1999a) Restricted maximum likelihood from symmetry. Statistical Research Report No. 14/1999. Department of Mathematics, University of Oslo. Available at http://folk.uio.no/ingeh/publ.html.

Helland, I.S. (2000). Some theoretical aspects of partial least squares regression. To appear in *Chemometrics and Intelligent Laboratory Systems.*

Helland, I.S. (2001). Reduction of regression models under symmetry. Invited contribution to: Viana, M. and D. Richards [Ed.] *Algebraic Methods in Statistics.* Contemporary Mathematics Series of the American Mathematical Society.

Helland, I.S. (2002a). Statistical inference under a fixed symmetry group. Preprint, http://www.math.uio.no/ingeh/publ.html.

Helland, I.S. (2002b). Discussion of McCullagh, P. (2002) What is a statistical model? *Ann. Statistics* **30**, 1225-1310.

Helland, I.S. (2003a). Extended statistical modelling under symmetry: The link towards quantum mechanics. Submitted. Available on http://www.math.uio.no/ingeh/publ.html.

Helland, I.S. (2003b). Quantum theory as a statistical theory under symmetry and complementarity. Submitted. Available on http://www.math.uio.no/ingeh/publ.html.

Hora, R.B. and R.J. Buehler (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* **37**, 643-656.

Kariya, T. (1989). Equivariant estimation in a model with an ancillary statistic. *Ann. Statist.* **17**, 920-928.

Kass, R.E. and L. Wasserman (1996). The selection of prior distributions by formal rules. *J. Amer. Stat. Ass.* **91**, 1343-1370.

Lehmann, E.L.(1959). *Testing Statistical Hypotheses.* Wiley, New York.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation.* Springer, New York.

McCullagh, P. (1996). Linear models, vector spaces, and residual likelihood. In: *Modelling Longitudinal and Spatially Correlated Data.* Editors: *Gregoire et al.* . Springer Lecture Notes No. 122, pp. 1-10.

Nachbin, L. (1965). *The Haar Integral,* Van Nostrand, Princeton, N.J..

Press, S. J. (2003). Subjective and Objective Bayesian Statistics. Wiley.

Schweder, T. and N.L. Hjort (2003). Frequentist Analogues of Priors and Posteriors. In: Stigum, B.P.: *Econometrics and the Philosophy of Economics. Theory-Data Confrontations in Economics.* Princeton University Press.

Stein, C. (1959). The admissibility of Pitman's estimator of a single location parameter. *Ann. Math. Statist.* **30**, 970-979.

Stein, C. (1965). Approximation of improper prior measures by prior probability measures. In: J. Neyman and L.M. LeCam (*eds.*) *Bernoulli, 1713; Bayes, 1763; Laplace, 1813*, 217-240. Springer-Verlag, Berlin.

Stone, M. (1965). Right Haar measures of convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **36**, 440-453.

von Neuman, J. (1932). *Matematische Grundlagen der Quantenmechanik.* Springer, Berlin.

Zacks, S. (1971). *The Theory of Statistical Inference.* Wiley, N.Y.

Résumé: On suppose q'un groupe de transformations s'exécute à l'espace de données et induit un groupe de transformations à l'espace de paramètre. Une classe de fonctions paramètric importante - les sous-paramètres permissibles - est intruit. Régions de credibilites (avec mesure invariante droite comme prior) et régions de confidence sommes démontrées d'être égal quand génerées par des sous-paramétres permissibles. L'effet du choix des groupes des transformations est illustrée par des exemples. On montre qu'il exist des relations contre methodes de prediction dans chemometrie, contre des design des expériences et contre de physique quantique.

Inge S. Helland
Department of Mathematics, University of Oslo
P.O.Box 1053 Blindern, N-0316 Oslo, Norway
E-mail: ingeh@math.uio.no